情報と文化

——テキスト・データベースの現状と展望——

長瀬真理

はじめに

高度情報化社会の担い手であるコンピュータの発展には目ざましいものがあり，ひと昔前の“数値”処理から“文字”，“画像”，“音声”といった文化と関連の深い領域にまで拡大している。その意味で，“文化”もすでにコンピュータによる，処理・加工・伝達可能な“情報”であると言えよう。なかでも情報の処理及び検索といった実際面の作業が，論理に裏付けされた言語，具体的には様々な人工言語で書かれたプログラムによる制御に基づくことから，文字情報の研究・発展の重要性はますます増大すると考えられる。

これに応じて、これまで理科系の学問に比して遅れていた人文系の研究にもコンピュータが徐々に導入されるようになった。しかし既に蓄積されつつあるデータや情報の評価，あるいは具体的な利用はもとより，それ以前の“一体何を情報やデータとすべきか”，といった基本的な事柄については，これまで充分に論議されていたわけではない。とりわけ文科系の学問における“情報化”の問題の究明，理論体系化，その応用を対象とする個々の学問分野の確立は非常に遅れている。本稿では，文字情報の代表ともいえるテキスト・データベースを取り上げ，それに関連する諸問題を検討しながら情報と文化の関わりについて考察する。又その作成の根幹に影響を与えるのみならず，図書館，出版，印刷等への多大な影響が予想される文字情報の入力標準化の動き(SGML)にもふれる。論述は以下の順に行われる。

1) テキスト・データベースの開発
2) テキスト・データベースの利用

3) 日英対照「源氏物語」テキスト・データベース

4) SGML

5) 将来の展望

1. テキスト・データベースの開発

人文科学系の研究にもコンピュータが利用されるようになった大きな理由の一つに，コンピュータの文字の処理技術の発展に加えて，容量が増大し，大量の資料やテキストを扱うことが可能になったことがあげられる。そこで盛んになったのが欧米を中心に始まった，文学や哲学等の古典を全体を丸ごと入力してしまうテキスト・データーベースの出現である。テキスト・データベースとは収集，整理，加工等，常時必要に応じてコンピュータ処理が可能なデータ・ファイルのことである。今日の光学読み取り装置の出現はこの作業の効率を高め，欧米のほとんどの国に大学や研究所を中心にしたアーカイヴができている。なかでも最も進んでいる Oxford University Computing Service (OUCS) では，テキスト・データベースの様々な利用法が検討され，積極的に研究や教育に取り入れられている。

1.1 テキスト・データベースの形式

テキスト・データベースの入力形式にはいくつかの典型的な形式がある。代表的なものとしては，固定長形式，COCOA 形式，があげられるが，現在は SGML（後述）の台頭と共に COCOA 形式の SGML への変換が進められている。

両形式の大きな違いは後からの解析に利用される参照部の宣言の処理方法にある。固定長では各行の頭に頁，行，章などの情報が付加されており，テキストのポジションがわかりやすい反面，その分入力の容量が大きくなる。又，参照部の情報が固定されているため，後から新たな参照事項を付加するのが困難である。

表 1 と表 4 はそれぞれ固定長形式と COCOA 形式の例である。前者はカリフォルニア大学の付属機関である Thesaurus Linguae Graecae (TLG) 作

表1. TLG作成のテキスト・データベース. プラトン『エウチプロン』冒頭部

A0059001; X2; YA; Z1..._=*E*U*Q.= *T4I NE4WTEPON, 25W *S4 WKPATE@, G4EGONEN, 34OTI S6 U T6A@ 2EN
A0059001; X2; YA; Z2..*LUKE4IWJ KATALIR6WN DIATPIB6A@ 2ENQ4 ADE N5UN DIATP4IBEI@ REP 6I T6HN
A0059001; X2; YA; Z3..TO5U BASIL4EW@ STO4AN; O2U G4AP ROU KA 6I SO4I GE D4IKH TI@ 02 5USA
A0059001; X2; YA; Z4..TUGX4ANEI RP60@ T60N BASIL4EA 34WSREP 2 EMO4I.
A0059001; X2; YA; Z5..._=*S*W.= *24UTOI D6H 2*AQHNA5IO4I GE, 25W *E2UQ4UFPWN, D4IKH N'A2UT6HN
A0059001; X2; YA; Z6..KALO5USIN 2ALL6A GPAF4HN.
A0059001; X2; YB; Z1..._=*E*U*Q.= *T4I F4HJ@; GPAF6HN S4E TI@, 3W @ 24EOIKE, G4EGPART AI: 02U
A0059001; X2; YB; Z2..G6AP 2EKE5IN40 GE KATAGN4WSOMAI, 3W@ S 6U 34ETEPON.
A0059001; X2; YB; Z3..._=*S*W.= *02U G6AP 025UN.
A0059001; X2; YB; Z4..._=*E*U*Q.= 2*ALL6A S6E 24ALLO@;
A0059001; X2; YB; Z5..._=*S*W.= *R4ANU GE.
A0059001; X2; YB; Z6..._=*E*U*Q.= *T4I@ 035UTO@;

成のギリシャ語のテキスト・データベースである。固定長形式の代表的な例であり，ギリシャ語のアルファベットはロマン文字に，気息記号やアクセント記号，大文字を示すサイン等は数字に置き換えられている。

一方COCOA形式では，処理対象となる本文と別に，表題，ページ，版，章，著者，登場人物などの情報をあらかじめ参照部として宣言しておく。そうすることにより，版，ページ，登場人物によって検索したり，並べ変えることが可能になる。参照部以外は，ほとんどオリジナルの本の形式と変わらないという利点がある。なおこのタイプの詳細は3章の「源氏物語」テキスト・データベースにおいて解説する。

1.2 サービス

テキスト・データーベースを利用を希望する研究者は，直接ネットワークを通して，あるいは磁気テープやフロッピィ・ディスクなど希望のメディアで安価にサービスを受けることが出来る。又最近はシェイクスピアのように

商品として購入することも可能になってきた。

OUCSの場合は，サービスされているテキスト・データベースのリストは小冊子の印刷物や電子メール等で知ることができる。それには著者名，タイトル，言語，容量のほか，利用条件についての情報が付いている[1)]。

1.3 問題点

以下簡単にテキスト・データベースのかかえる問題点について述べる。すべて未だ解決されておらず，今後の動向に充分注意する必要がある。

1.3.1 テキストの選定

テキストを入力する際に最も重要なことは，バージョンの選定である。出来れば入力には時間もコストもかかること，更に多くの人々に公開することを前提に，汎用性のあるテキストを選定すべきである。後述するコピーライト（著作権）の問題も関係してくるが，安易に著作権がクリアできることを採用の基準とすべきではない。データベースのメリットの一つが国際間での共有にあることを考慮するならば，多くの研究者に権威あるテキストとして受け入れられているものを原本とすることが望ましい。

1.3.2 テキスト・データベースの評価

光学読み取り装置の発達と共に，テキスト・データベースの最初の入力に機械を利用することが多くなったが，印刷の鮮明度や書体，紙質など条件により読み取り率は異なる。そのため校正が重要な作業になっており，テキスト・データベースの善し悪しは何度校正がなされたかに依存するといっても過言ではない。又，バージョン・アップのサービスの有無も評価の大きな要因である。先のTLGは5年ごとに更新しており，何度校正を行ったかが明記してある[2)]。

1.3.3 コピーライト

個人でサービスを行うことは実質的に費用もかさみ困難なため，多くの場合国際的な機関に供託しサービスを依頼することが多い。しかし供託を受けた側がそれを無断で第三者にコピーすることは許されない。なぜならテキスト・データベースの作成と同時に制作者にコピーライトが生じ，制作者の権

利を保護しなければならないからである。

例えばOUCSでは，厳しい指定条件を作成しており，テキスト・データベースの利用者は全員この基準内容を示した書類(Condition of Use)に署名しなければならない。書類にはテキスト・データベースが学術研究にのみ使用されることの保証や複製の禁止が明記されている。

OUCS自身が作成したテキストの大部分は古典であり，著者の死後50年以上経過したものがほとんどである。その意味では既にパブリック・ドメインになっており，本をテキスト・データベースにする段階でコピーライトの問題は生じていない。

しかし，日本では一般に出版社が作家の著作権を買い取ることが少ないことから，出版社の中には編集著作権の他に複製権を主張する場合が少なくない。そのため，校訂本の全文を光学読み取り装置等で入力する場合，出版社に事前に了解を得る必要がある。

いまのところ，ソフトでは色々問題が起こっているが，テキスト・データベースに関しては，コピーライトが問題になった例はない。外国ではこういったトラブル専門の職業としてLiterary Executerがある。

現在ではテキスト・データベースがある方が本の売行きも伸びることから，制作に協力的な出版者が増えている。

2. テキスト・データベースの利用

次にここでコンピュータ処理が可能になったテキスト・データベースをどのように利用するかが問題となってくる。テキスト・データベースを電子化テキスト等と呼ぶ場合があるため，書籍との類比から，本の方が読みやすいとか持ち運びに便利だ，といった非難を聞くことがある。しかしコンピュータによる“読書”はこれまでの読み，あるいは研究スタイルの延長上にはない。テキスト・データベースはプログラムと呼ばれるコンピュータ上で走る道具を使うことによって初めてその威力を発揮する。

2.1 機械的な処理の肩代り

道具としてのコンピュータは人工言語の指示に従って仕事を行う。具体的には論理的に書かれた一連のプログラムによって制御されている。理科系の学問では，その主体となる数式計算において，数の比較や四則演算を始めとする様々な繰り返し演算をコンピュータに肩代りさせることが可能になった。コンピュータの高速かつ正確な演算は，天文学的な数字であるロケットの走行距離などを瞬時に計算することを可能にした。

同様に文科系の研究のプロセスのなかにも機械処理が可能な繰り返し作業があり，その作業をコンピュータに肩代りすることができる。コンピュータが複雑な数式の計算にもたらした革命は文科系の研究にも恩恵を施したのである。

例えば研究者が文学や哲学のテキストを何度も読む場合，常に最初から読んだりはしない。むしろ後のインディックスあたりから先に見ることの方が多い。うろ覚えだった文章やフレーズの確認，重要と思われる語や，熟語の出典箇所を調べたり，特定の語の用例を幾つも捜したりする。あるいは論文のテーマになりそうなキーワードを捜したり，特定の話者の性格や，心理描写だけを追ってみたりする。そして，それを整理するためにノートやカードを作る。これらのカード類は研究の深まりとともに増大し決して減ることはない。最終的には，テキストの一語一語が有意味になり各単語にカードが必要になってくるかもしれない。しかし，こういった語句の検索や語彙の用例捜しに代表される機械的な作業の大部分は，コンピュータが最も得意とするものである。とりわけ研究者の関心が，長いテキストの全体にある場合，コンピュータの有効性は高まる。頻度の高い語彙の用例をテキスト全体にわたって調査するのは手作業では困難で時間がかかる。しかしコンピュータを使うならば高速で正確に上記の作業を行ってくれる。又インディクス機能を使ってテキスト全体の自分用の辞書を作成することも可能である。あるいは語尾からの検索機能（リバース・インデックス）を使えば詩の韻律の抽出なども容易である。

研究者はむしろ最初の，どういった語に注目すべきか，といった創造的な作業にのみ集中すればよい。即ち，コンピュータの利用は文科系の研究方法を客観的な吟味を可能にしたともいえる。研究のプロセスに含まれる曖昧さを除去し，どこまでが機械的な作業であり，どこからが創造的な作業であるかを明確にすることができるようになった。もっとも現在のところ，必ずしも文科系の研究者の要望を念頭において作られたソフトは少ないことから，まだまだ研究の効率化を図る余地は充分あると言える。

一方コンピュータの処理機能はロジックにより限定されており，いくつかの機能を組み合わせたとしてもそのコンビネーションには限りがある。しかも人間の頭脳というものは必ずしも論理的な働きのみが支配している訳ではない。その機能は未だ解明されておらず，創造はしばしば直感や飛躍の産物である。ある意味で現在行われている，曖昧性を考慮したファジイ・コンピュータも含めて人工知能やロボットの研究は，どこまで人間の創造的機能に迫れるか，というテーマに収斂されるかもしれない。このような観点に立つならば文科系の研究方法の吟味は，これまでのコンピュータの限界内での技術的な可能性に着目していた理科系主体の発想の吟味にも通用する部分があると言えなくもない。即ちこれからは限界内での技術の進歩や利用の拡大と同時に，限界からの超越や飛躍を促す発想もコンピュータの発展には必要なのである。むしろ今後の文科系の研究用ソフト開発を促進する上でも，論理に支配されない発想が新しい視点を提供する可能性が高いかもしれない。今後の人文系でのコンピュータ導入の拡大が，人間の創造力究明の鍵を握っているとも言える。

文科系の研究へのコンピュータ利用との関連で忘れてならないのは，テキスト・データベースの開発で言われたメリットが研究成果に関しても当てはまる，という点である。理科系の学問とは違って，結果の即効性は重要視されないが，共有化という点は大いに今後利用されるべきである。ネットワーク等の利用は多くの人々に研究成果へのアクセスを可能にする。情報となった様々な研究成果や研究方法はこれまでとは違って各国の研究者に共有され

るであろう。その結果，学際的さらには国際的な研究の広がりや交流を可能にする。

現在文科系の研究で利用されるプログラムは基本的には検索・並べ換え・頻度抽出等の作業が中心である。これらの機能によってインデックス・ワードリスト，KWIC に代表されるようなコンコルダンスを容易に作成することが出来る。

2.2 著者推定

コンピュータの利用によって最も成果が期待されているのが著者推定等に代表される文体研究である。

例えば，1980 年 7 月 6 日，ロンドンの代表的な日曜紙「ザ・オブザーバー」はその第一面で“コンピュータが埋もれていたシェークスピアの作品を見つけだした”と報じた。実際はコンピュータではなく，ベーシングストーク・テクニカル・カレッジの T. メリアム氏がコンピュータを使って「“サー・トマス・モア”の台本」と呼ばれる作品を鑑定したのである。古来このテキストばかりでなく，シェイクスピアの著作の真偽に関しては様々な説がある。シェイクスピア＝ベーコン説を筆頭に，クリストファー・マーロー，妻のアン・ハサウェイ，あげくのはてはエリザベス女王の名まで挙がっている。

メリアム氏の使った手法はエジンバラ大学の A. Q. モートン教授の開発したもので，現在では文体統計学 (Stylostatistics) あるいは文体計量学 (Stylometrics) と呼ばれる新しい分野で盛んに使われている。教授は，“作家には言葉の使用に関して，それぞれ無意識に使用する独特の癖があり，それらは他人には真似できないもので，いわば文体上の指紋である”という。

メリアム氏は比例対（a と an, all と any といった文中で似たような働きをする語の対）反対語対（with と without や can と cannot 等）の頻度や語連結（to be, of course, as if, and the 等）の使用頻度や，it, a, the 等の文頭での使用率，珍しい単語（使用頻度の低い外来語や専門用語等）の使用範囲など様々な観点から作家の“指紋”を調査した。

その結果，上記の「台本」は他のシェイクスピアの真作とされる29作品と41項目の癖の内40個が一致し，真作の仲間入りをした。

最近も戯曲ではなく，“私は死ぬべきか，逃げるべきか(Shall I die?, Shall I fly)”で始まる90行の恋愛詩の発見にコンピュータが一役買った。オックスフォード大学出版局の編集者のG. テーラー氏は，「新シェイクスピア全集」の編集作業中，ボードレアン図書館から借りだした膨大なカタログの中からこの詩を見つけだした。署名がシェイクスピアとなっていたばかりか，コンピュータで作成された用語辞典を使って語彙を調べた結果，シェイクスピアが30才位の時に書いた「ロミオとジュリエット」と類似していることから，シェイクスピアの真作との確証を得たという。これは米国でも評判になり，「ニューヨーク・タイムズ」や雑誌「タイム」などで大きく取り上げられた。

メリアム氏やテーラー氏に対する反響は大きく賛否両論が続出したが，今ではシェイクスピア研究にコンピュータはなくてはならない道具となり，この種の研究論文が「Journal for Literary and Linguistic Computing」や「Computer and Humanities」等の欧米の学会誌や論壇をにぎわせている。

2.3 問題点

コンピュータの利用は始まったばかりであり，今後多方面で利用されると思うが，まだまだ解決されなければならない問題も多い。そのため利用に際しては充分その限界を知っておく必要がある。例えば，コンピュータはアルファベットが一緒でも意味が異なる同音異義語や，ローマ数字とアルファベットの区別，固有名と一般語の区別等をすることはできない。又語幹の異なる語の形態素分析や意味論等に利用できるソフトは開発されてはいない。語形をすべて指定したり，あるいは特定の語を排除したりすることが必要になる。又文脈と一緒に取り出しても最終的には人間が判断しなければならないケースも多々あり，かえって面倒な場合もある。頻度の多い単語にしてもすべて抽出してしまうため返って重要なケースの選択が煩雑になる恐れもある。コンピュータの進歩にはめざましいものがあり，限界は次から次ぎへと

打ち破られつつあるが，その発展を追うのも大変であるばかりか，費用もかさむ。その意味ではそれぞれの関心や得意な分野を生かしたプロジェクト式の研究が望ましく，今後理科系のようなグループ研究が増えるのではないかと思われる。

3. 日英対照「源氏物語」テキスト・データベースの作成

ここで実際に筆者が作成した日英対照「源氏物語」テキスト・データベースを取り上げ，具体的なテキスト・データベースの制作や利用について解説する。

日本語版は1987年小学館発行の「源氏物語」(阿部秋生，秋山虔，今井源衛校正注訳)，英語版はE. G. サイデンスティッカー訳「The Tale of Genji」タトル社を原本として使用した。

1990年4月よりOUCSからサービスされており，米国や英国の研究者によって利用されている。

3.1 入力方法

入力は実験も兼ねて光学読み取り装置を利用した。日本語には富士電気総設の富士OCRシステムを採用し，入力を委託した[3)]。文字認識のアルゴリズムは原稿をイメージ・データとして取り込み，文字線の境界に波を伝播させ，その波頭から文字の高次の特徴を抽出するという手法を採用している。英語版の入力は，既に各国で使われ定評のあるKurzweil 4000を使用した。

3.2 検索ソフトについて

2章でも述べたように折角テキスト・データベースを作成しても検索ソフトがなければ，利用の効果は余り期待できない。今回のプロジェクトではソフト開発は含まれていないため，既存の検索ソフトの利用を想定しなければならない。

都合の良いことに，供託先のOUCSでは様々なテキスト・データベースを作成すると同時に，政府からの基金援助を得て，パッケージ・プログラムの開発も盛んに行っている。そのなかに，オックスフォード・コンコルダン

ス・プログラム（Oxford Concordance Program：略 OCP）と呼ばれる文章解析プログラムがあり，供託されたテキスト・データベースの整備やスペル・チェクにも利用されている。国内での公開の引き受け先である東京大学大型計算機センターにも，1987 年より導入されている。

又，パソコン版 Micro-OCP が作られ，日本語の文章解析も出来る為，今回のプロジェクトのように日本語・英語のテキスト・データベースの処理には最適なソフトであると考えられる事から，OCP を解析プログラムとして採用した。

一般に多くの機械翻訳や文章解析用ソフトは，テキスト・データの分かち書き，キーワードや読みの付与，形態素分析を前処理として心要とする。しかし Micro-OCP の場合はその必要がなく，ベタ書きのテキスト・データでもそのまま処理できるという利点がある。

3.3 入力形式

入力形式については 1 章で述べられたように様々な方法があるが，英文，和文共に検索システムとして OUCS が開発した Micro-OCP を使用を予定しているため，COCOA 形式を採用した。この形式は，OCP が最も容易に取扱うテキスト・データ形式で，アトラス研究所時代から使われている。ちなみに COCOA の名前は word COunt COncordance generation on Atlas の大文字部分から採られている。

表 2 と表 3 はそれぞれ COCOA 形式に則った日本語及び英語の参照部の定義であり，表 4 は実際のテキスト・データベースである。

日本語の原本は縦書きで，上段に語彙の註，中段に本文，下段は現代語訳と三段に分かれている。又，所々に挿絵が挿入されている。

テキスト・データベースでは本文のみが横書きで入力され，註，現代語訳，挿絵は除かれた。又本文中に現れる註番号や振り仮名も取り除かれた。なお行の切れ目は原本と同じである。

参照部として今回採用された情報は，作者名 (W)，表題 (T)，巻数 (V)，章 (C)，英語の章 (E)，小見出し (S)，小見出しに対応する英文ページ (F)，日本文

ページ(P)，登場人物(A)である。これらの情報はカテゴリーと呼ばれ，それぞれの後ろの( )内に示されたようなアルファベットの1文字変数として鍵括弧(〈〉)でくくって宣言される。

本の場合と違って英文との対応を示す幾つかの情報が付加されている。先ず章に関しては，日本語も英語も各々54章に分かれているため相互に対応する。しかし日本語の章見出しには数字が無い。その為英文から日本文への参照を容易にするため，あえて英文の章番号を挿入した。更に細かいレベルでの対応については，英文には無いが，日本文では章が又いくつかに分けられ，番号順に小見出しが付けられている。この点に着目し，日本文の小見出しに対応する部分についてのみ英文のページを挿入することにした。その為小見出しの冒頭の部分は相互に対応させることが出来るようになった。

更に本では，総てではないが会話部分について，話者の情報が挿入されている。それを生かすために，登場人物として話者マークのカテゴリー〈A〉を

表2. 日本語参照部の定義

〈W紫　式部〉Writer's name: 作家
〈T源氏物語〉Title's name: 題名
〈V No.〉Japanese volume (1-6): 日本語の巻
〈C きりつぼ〉Japanese Title of Chapter: 日本語の章
〈E No.〉English Chapter (1-54): 英語の頁
〈S No.〉Japanese subheading: 日本語の小見出し
〈F No.〉Page references of the English texts: 英語の頁（英文箇所に対応）
〈P No.〉Page references of the Japanese texts: 日本語の頁
〈A 帝〉Names of Speaker (Actor or Actress): 話者

表3. 英語の参照部

〈W Murasaki Shikibu〉Writer's name: 作家名
〈T The Tale of Genji〉Title's name: 題名
〈K No.〉Japanese volume (1-6) 日本語の章
〈C No.〉Chapter (1-54): 英語の章
〈N No.〉Page references of the Japanese subheading: 日本語の小見出し（日本語と対応）
〈P No.〉Page references of the English translation: 英語の頁

表 4. 日英対照「源氏物語」

〈W 紫式部〉
〈T 源氏物語〉
〈V 1〉
〈C きりつぼ〉
〈E 1〉{The Paulownia Court}
〈S 1〉{桐壺更衣に帝の御おぼえまばゆし
〈F 3〉
〈P 93〉
いづれの御時にか，女御更衣あまたさぶ
はあらぬが，すぐれて時めきたまふあり
御方々，めざましきものにおとしめそね
は，ましてやすからず。朝夕の宮仕につけ
りにやありけん，いとあつしくなりゆき，
あはれなるものに思ほして，人のそしり
き御もてなしなり。上達部上人なども，あ
ぼえなり。唐土にも，かかる事の起りに
〈P 94〉
れと，やうやう，天の下にも，あぢきなう
引き出でつべくなりゆくに，いとはした
たぐひなきを頼みにてまじらひたまふ。
父の大納言は亡くなりて，母北の方な
し，さしあたりて世のおぼえはなやかな
ももてなしたまひけれど，取りたてて，に
ほ拠りどころなく心細げなり。
〈S 2〉{更衣に皇子誕生，方々の憎しみつ
〈F 3〉
前の世にも，御契りや深かりけん，世に
ぬ。いつしかと心もとながらせたまひて，
の御容貌なり。
一の皇子は，右大臣の女御の御腹にて，
かしづききこゆれど，この御にほ
〈P 95〉
ひには並びたまふべくもあらざりければ，
君をば，私ものに思ほしかしづきたまふ
はじめよりおしなべての上宮仕したま
となく，上衆めかしけれど，わりなくまつ
りをり，

〈W Murasaki Shikibu〉{Translated by
〈T The Tale of Genji〉
〈K 1〉{Japanese Volume}
〈C 1〉{The Paulownia Court}
〈N 1〉{Japanese Sub-chapter
〈P 3〉
In a certain reign there was a lady not o
more than any of the others. The grand
a presumptuous upstart, and lesser ladi
she did offended someone. Probably
seriously ill and came to spend more tin
pity and affection quite passed bounds.
courtiers might say, he behaved as if in
His court looked with very great
infatuation. In China just such an unrea
an emperor and had spread turmoil thr
the example of Yang Kuei-fei was the or
She survived despite her troubles, w
of love. Her father, a ran councillor,
old-fashioned lady of good lineage, was
for her than for ladies who with paterna
The mother was attentive to the smalles
there was a limit to what she could do. T
strong backing, and each time a new in
〈N 2〉
It may have been because of a bor
〈P 4〉
emperor a beautiful son, a jewel beyond
impatience to see the child, still with
earliest day possible, he was brought to
marvelous babe. The emperor's eldest
the Right. The world assumed that with
be named crown prince; but the new c
occasions the emperor continued to fav
private treasure, so to speak, on which
The mother was not of such a lov
personal needs. In the general view she
on having her always beside him, howe
or other entertainment be would require
of them would sleep late, and even afte
Because of his unreasonable demands
immoderate habits out of keeping with
With the birth of the son, it becam
favorite. The mother of the eldest so
manage carefully, she might see the nev
come to court before the emperor's othe
the others, and she had borne severa
complaining might trouble and annoy h
ignore.
Though the mother of the new s

導入した。その結果，特定の話者の会話部分のみを処理対象に指定することが可能になった。

原本では挿入のような取扱になっている小見出しの数字に続くタイトル状の解説は {} でくくり，コメント（注釈）とした。又英文の各章のタイトルも注釈扱いとした。OCP は注釈を本文とみなさないが，こうしておけば解析時に処理対象として宣言することも，あるいは逆に削除することも容易である。

英文の場合も，入力されたのは本文のみで，脚注や各ページの先頭中央のタイトル及び挿絵は省かれた。

参照部は，一見して明らかな通り日本語より短くなっている。両者の大きな違いは，日本語の解説時にも触れられたが，小見出しが無いことにある。もう一つ，英文では会話部分はダブルコーテーション（”）で囲まれているが，一体誰が話しているか，と云った情報が原本にはない。その為話者マークのカテゴリーもない。

日本語と英語の対照として日本語の巻の情報を入れたのは，日本語は全部で 6 巻に分かれており，各巻 1 からページ付けが行われていることによる。他方英文は 2 巻に分かれているが両者には通しページが付いている。

なお既に解説したように，日本文の小見出しの冒頭部のページが参照部として入力されている。番号順にはなっているが，小見出しに続く節が長い場合には，後半部の対応に大きなずれが生じている。

3.4 日本語の構成

日本語の本文を構成する基本文字は平仮名と漢字であり，句切り文字は句読点，括弧（「,」）（『,』）である。只，JIS の第 2 水準にもない漢字が若干あり，その場合は類似のものを代入し，暫定文字としてその文字の後ろにアスタリスク（*）を挿入した。

次行に続く漢字語句の場合は行末にハイフンを挿入した。漢字のみとしたのは，日本語版の Micro-OCP を用いる際，平仮名を処理上句切り文字として扱うよう指定したため，たとえハイフンを挿入しても，コマンド・レベル

で指定しない限り，平仮名の検索は一切しないからである。又現実問題として，日本語の表記法の解釈は様々で語彙をどこで区切るかは必ずしも専門家の間でも意見の一致がみられていない。むしろ各人の自由な解釈にまかせ，それぞれ表記法のもとで語彙の形が設定されるほうが望ましいと判断した。

ハイフンで一番困ったのは“御”の処理である“御簾”，“御覧ず”，や“御息所”のように他の漢字と切り離せない場合もあれば，“御心”，“御気色”，“御消息”のように切り離して使われる場合もある。今回はいずれの場合も区別せず，2行に亙る時は機械的にハイフンを付ける処理をした。

3.5 英語の構成

英文の基本構成文字はアルファベットであり，その他句切り文字として句読点及びコロンが使われている。会話箇所の開始と終了にはダブルコーテーション(")，行にまたがる単語の継続記号としてハイフン(-)を，更にアポストロフィ(')が使われ，これらは原文とほぼ同様である。

3.6 Micro-OCP による解析例

表5はMicro-OCPによる分析例である。このソフトはワード・リスト，インデックス，並びにキーワードを文脈と共に抽出する用語索引を作成する機能を持つが，表の例は本文一行を伴ったKWICの出力結果である。

本居宣長以来「源氏物語」の基調語として有名な“あはれ”を1章の「桐壺」から抽出している。「源氏物語」全体では1024箇所に使用されており，これら全ての出力も容易である．E. G. サイデンステイッカーの翻訳と対応する大まかなページ参照部を利用するならば両テキストの比較研究も可能である。

その他単語検索にワイルド・カードの機能を使えば，語尾変化をする語や敬語の語彙研究などにも応用できる。

又2章でも論じたが真偽問題の解明にコンピュータは盛んに利用されている。「源氏物語」にも同様の問題がある。いわゆる“宇治10帖問題”といわれているもので，最後の10帖と他の44帖との関係について長年議論がなされている。

コンピュータを使ったものから，使わないものまで様々な研究がなされているが未だ結論はでていないようである。

今回の英語版の原本の翻訳者であるサイデンステイッカーは，先のシェイクスピア＝ベイコン説にならって，この問題を“Genji-Baconian Theory”あるいは“Genji Baconians”と命名し興味深い議論を展開している。氏は，42帖の「匂宮」の冒頭で源氏が亡くなるところで物語は終り，42とそれに続く43の「紅梅」，44の「竹河」の3帖を変わり目，あるいは経過の部分とし，45以降から新しい物語がはじまると主張している[4]。

3.7 OUCSへの移植

テキスト容量は圧縮ソフト等を使用すれば1 Mbyte程であるが，利用しやすいように英文，日本文共54ファイルとしたため最終的にはそれぞれ3 Mbyteになった。OUCSへの供託に際しては，日本語の部分をどのようなコードで入力するかが大きな問題となった。

日本語と英語等ヨーロッパ系の文字の大きな違いは，日本語の文字がアルファベットに比して非常に沢山の種類があることにある。そのため一般に日本国内では，英文字のように1 byteの半角文字ではなく漢字や仮名は全角の2 byte（16ビット）で表示している。

又筆者の使用しているNECパソコンのOSである日本語版MS-DOSとIBMパソコンのOSであるIBM-DOS，いわゆるPC-DOSとは細かい部分で異なる。特にIBMは日本語表示の為のハードを持たない。更に困ったことにはFDにはフォーマットの形式が色々あるのに加えて，日本で売り出されているIBM機種は日本語ワープロを搭載しているため欧米のものとは互換性が無い。その他，細かいことになるが高密度の外部ディスクの容量もIBMと日本国内で使用されているものとでは異なる。

そのためフロッピィでの提供は2DDディスクでもうまくいかなかった。結局最終的には日本語文字識別子の制御コードさえ読めれば移植が容易なISO（国際標準規格）に則った7ビットコードに変換し磁気テープで供託した。

表5. Micro-OCPによるKWICの出力例

あはれ　12

1きり 1 3 93 9　るを、いよいよあかずあはれなるものに思ほして、人の
1きり 2 3 96 12 思ひわびたるを、いとどあはれと御覧じて、後涼殿にもと
1きり 4 5 98 5 、いたう面痩せて、いとあはれとものを思ひしみながら、
1きり 5 6 100 9　きわざなるを、ましてあはれに言
1きり 6 6 101 13 みたまひしか、人がらのあはれに、情ありし御心を、上の
1きり 8 7 103 4 引き入るるより、けはひあはれなり。やもめ住みなれど、
1きり 9 10 109 7　せたまはざりけると、あはれに見たてまつる。御前の壺
1きり 9 10 109 14 ありさま問はせたまふ。あはれなりつること、忍びやかに
1きり 9 10 110 13　はせて、いとあはれに思
1きり 12 14 116 11　りたるに、皇子もいとあはれなる句を作りたまへるを、
1きり 13 15 119 9　思し慰むやうなるも、あはれ
1きり 14 16 120 2　を、若き御心地にいとあはれと思ひきこえたまいて、常

日本文の行
日本文のページ
英文のページ（日本文の小見出しに対応）
日本文小見出し番号（英文のページに対応）
テキスト名
日本文巻数

統計量

TOTAL WORDS READ = 1927
TOTAL WORDS SELECTED = 1927
TOTAL WORDS PICKED = 12
TOTAL WORDS SAMPLED = 12
TOTAL WORDS KEPT = 12
TOTAL VOCABULARY = 1

しかし実際OUCSがメインフレームとして使っていたのはVAXマシンであった。このマシンのOSはVMSという種類で，日本語に関してはATTコード（DECコードともいわれる）を採用しており，残念ながらISOコードでは判読不能である。なおOUCSにはUNIXが走るSUNのワーク・ステーションも入ってはいたがテキスト・アーカイヴ部門ではVAXが使われており，SUNはまだ完全な実用段階にはなっていなかった。その為JISコードに変換をしなければならなくなったわけだが，好運にもATTコードに変換しなくてもそのまま読み込むことのできるMTR (magnetic tape reader)というソフトを中国語のテキスト・データベースの移植作業をしている担当者から入手することができた。なお今後のUNIXの普及とパソコンへのダウン・ロード作業を考慮し，急遽アーカイヴ部門にもSUNを接続してもらい，出来上がったファイルをE-mailを使ってSUNのワーク・ステーションへ移した。

オックスフォードでも文科系ユーザーはパソコン主体に移りつつある。共同研究者の使っていた日本語ワード・プロセッサーはEW-Plusというソフトで，容量を少くするためパソコン専用のシフトJISを採用していた。その為今度はJISコードをワープロ用のいわゆる8ビットのシフト-JISにかえなければならなくなった。そこでSUNに蓄えられたファイルをパブリック・ドメインになっているNKF (network kanji filter) ソフトを使ってシフトJISに変換した。このソフトは，別名Three Way Converterとも呼ばれ，自動的に7ビットJISコード，シフトJISコード，EUC (Exteded Unix Code)コードの3種間の変換を行ってくれる便利なソフトで，MS-DOSバージョンもある。提供者はパソコン用のオブジェクトしか持っていなかった為，今後他の日本語テキスト・データベースの供託も予想されることから，日本からE-mailを使ってC言語で記述されたプログラム本体を送ってもらいオックスフォードに寄贈した。以上の作業を終えた段階で，日本でも普及している通信ソフトのKERMITを使ってSUNからIBMパソコンにファイルを落とした。

現在異なるコンピュータ間での情報交換を支障なく行う為に，ISO（国際標準化機関）で検討が進んでいる。即ち世界中のできるだけ多くの文字に適用する統一した文字符号を作成することを目指している。そこでは全世界の文字を表現するのに充分な容量を確保するために，一文字を32ビットで表記しようとしている。そうなれば10億以上の文字を区別できるようになる。しかし漢字だけをとっても，日本，中国，韓国はそれぞれ異なった方向で簡素化を行ってきている。まずは各国間での統一化をどうするかという問題が生じてこよう。又ISOの動きを牽制するような形でユニコードと呼ばれる16ビットによるコード化を検討するグループもある。

4. SGML

欧米では新たにTEI (Text Encoding Initiative) やCentre for Computing in the Humanities (Toronto University) の研究者グループが中心になって，SGML（Standard Generalized Markup Language, 標準一般化マーク付け言語）方式による付加価値の付いたテキスト・データベースの研究や実験的な作成が盛んに行われている。SGMLは文書の論理構造，例えば著者や表題等のデータ項目を後から項目別に検索できるように，タグとして挿入しながら作成する方式で，多面的な読みを可能にするばかりでなくハイパー・テキストの作成にも威力を発揮する。既にISOやECで，そのタグの標準化がビジネス・ドキュメントを中心に進められている。COCOA形式に似ていることから，オックスフォード・アーカイヴでサービスされている900以上のテキスト・データベースのSGMLへの変換も現在進行中である。OUCSが開発したOCPのバージョン・アップもSGML対応が第一課題になっている。SGMLは文書間の交換を容易にする為の一種のメタ言語であり，文書を構成する要素とその要素間の論理的な構造を記述する。個々の文書や文字の属性を定義しておけば，媒体や出力装置に依存することなく文書の内容のみを交換可能にする。属性情報には文章の階層的な論理構造を示す，“表題”，“見出し”，“節”や“脚注”を始め“倍角”や“イタリックス”

といった文字の指定等も含まれる。

既に米国の出版協会，国防省始め，EC も採用を決め，現在標準化の作業が進められている。特に共通語を決めなかった EC ではこの作業が急務となっている。1992 年の経済統合を目前にひかえ膨大な資料を各国の言語に翻訳しなければならず標準化は機械翻訳を促進するためにも重要な課題である。今後チェコスロバキア等東欧諸国の加入の可能性も高まり SGML への期待はますます大きくなりつつある。日本でも現在 SGML を JIS 規格にする動きがでている[5]。

今後「源氏物語」のテキスト・データベースも SGML 方式による作成が可能であり，それは専門家のみならず，学生や一般の人々にも，テキストの多面的な読みを可能にする。例えば，「源氏物語」では登場人物の呼び名や官職がしばしば変わる。主人公の源氏は，光，君，若，おとど，六条院，大将，大殿，院，男君など 30 以上の呼び名を持つ。又中国の文献からの引用も多い。そのたびごとに，後の付録の家系図などをいちいちひっくり返して調べながら読み進まなければならない。こういった項目をタグとして挿入し，参照部として別ファイルに家系図や外来語を入力しておけば，コンピュータの画面上で，それらの情報へ飛んだり，画面分割によって同一画面で調べることも容易になる。又呼び名の変遷を通してテキストの順番や書かれた年代の推定等の研究もできる。あるいは和歌，注釈のタグを付加することも可能であろう。そうすれば読者は和歌だけを取り出して研究することもできる。又現在，絵巻物が出版されている。絵巻と一緒に，今回削除した平安時代の服装や小物等の挿絵を一緒に画像データベースとして組み込むことも技術的には既に可能である。更に，源氏は優れた音楽演奏者で，宴の場面や楽器の演奏場面もよく出てくる。古代の楽器の奏でる音楽をミュージック・ソフトの技術で再現することも可能である。最早，文字と音と絵とを組み合わせたハイパー・テキストの作成も夢ではなくなっている。

5. 将来の展望

最後にテキスト・データベースおよび，研究面での利用の新たな可能性，並びに今後日本でも重要な課題になるであろう電子化辞典について述べる。

5.1 パラレル・テキスト・データベースの作成

日本研究が世界各地で盛んになり，もはや日本に行かなければ日本研究は出来ないという時代は終った。その意味で研究と教育の両方に役立つ，日本語と英語の相互対照が可能な，いわゆるパラレル・テキスト・データベースの需要は多い。我々が様々な外国文学研究で苦労するのと同様，海外の日本研究者や学生は日本語に苦労している。日本語と翻訳の両方の比較はもちろん，語彙の研究にもコンピュータは大いに威力を発揮するであろう。オックスフォード大学では，筆者の供託した「源氏物語」テキスト・データベースを使ったコンピュータによる“あはれ”の語についての日英比較研究のプロジェクトが発足している。残念ながらテキスト・データベースに対する理解や開発という点で，我国のこの分野での世界への貢献は経済の図式とは逆に輸入超過の傾向にある。その意味で今後より多くの日本文学や古典及びその翻訳書のテキスト・データベース作成が必要であろう。併せて日本国内にもテキスト・データベースを海外にサービスする機関の設置が急務となろう。

5.2 複数のテキストの比較研究の可能性

一つのテキスト・データベースを作成すると必ずその選択に不満を持つ研究者がでてくる。例えば「源氏物語」でも青表紙本や河内本，別本を始め，全ての校訂本や写本までもいれるべきだ，という意見は当然予想される。しかしこれは今日のコンピュータの能力を考えると不可能なことではない。現に英国では，写本研究や比較文献学，あるいは学生の教育用に，幾つもの版を入力したハイパー・テキストやソフトが開発されており，普通では入手しにくい貴重な文献に誰でもアクセスすることが出来るようになっている[6]。

5.3 電子化古語辞典の必要性

コンピュータによる古典研究の環境を整備するためには，内在資料としてのテキスト・データベースや検索ソフトの開発が必要であるばかりでなく，

様々な背景への理解を助ける為の外在的資料として，同時代に書かれた他の作品のテキスト・データベースが必要であると同時に，電子化古語辞典が必要不可欠である。既に海外では様々なテキスト・データベース作成と同時に，SGML 方式による辞書作成が盛んに行われており，OED (Oxford English Dictionary) の CD-ROM 版は大きな成果をあげている。これを使うと，語源がアラビア語あるいはヘブライ語の言語を十八世紀のテキスト中から捜す，あるいはディケンズのテキストから捜す，といった昔なら一人の研究者が何年もかかるような作業があっという間に出来る。辞書を一番使うのはいうまでもなく研究者である。このような辞書をみると末端のユーザーである文学研究者の意見がいかに良く反映されているかが解る。欧米では辞書の権威は学者・研究者の使用に足るものかどうかで決まる。質の高い辞書学 (lexicography) の長い伝統の蓄積と最先端のコンピュータによるデータベース作成技術である SGML の結合の見事な成功例といえるだろう。

日本でも将来この種の大規模な辞書作成のプロジェクトは確実に必要となるであろう。言葉は数字と違って曖昧である。重要な語や語句程，意味も広く，多様であり，沢山の用例研究を必要とする。日本文化の重要なキーワードである，“わび”，“さび” なども，時代の流れ，解釈により意味は多様化している。それらの語の用例が日本の古典の中から全て検索することが可能になれば，意味の変遷の研究に大いに役立つであろう。先に挙げた “あはれ” にしても，「源氏物語」だけでなく，「伊勢物語」や「古今集」等を網羅した電子化辞書が出来れば，研究者のみならず多くの人々が恩恵を受けるであろう。一方，消滅してしまって全く使われなくなった語句も多い。変化し崩れつつある日本語を保存するためにも古語辞典の電子化プロジェクトは必要であろう。

以上テキスト・データベースを中心に情報と文化に関連すると思われる様々な問題を検討してきた。利用やサービス面でのネット・ワークは整備されつつある。しかしサービスすべき情報自体の開発は遅れている。文化の根幹を成す古典的なテキスト・データベースの作成，それらの研究を支援する

ソフトの作成や教育用ハイパー・テキストの作成，そして何よりも電子化古語辞典の開発など課題は沢山あるが，今後関係者の貢献に期待したい。

謝　　辞

過去3年間に亘るテキスト・データベース作成のプロジェクト並びにオックスフォード大学での研究に対しては，社団法人「東京倶楽部」（理事長下田武三）の文化活動補助を受けることが出来た。テキスト作成に関して東京女子大学教授の秋山虔先生に大変御世話になった。又共同研究者である P. T. Harries 教授にはオックスフォード滞在中テキスト・データベースの利用に関して多くの示唆を受けた。ここに記して深く御礼申し上げる。

注

1) 利用条件は，U, X, A の三種類の記号によりランク付けされている。U は一般利用 (UNIVERSAL ACCESS) を意味し，このコードの付いたテキスト・データベースの利用者は先のコピーライトで述べられた利用条件の確認書に署名すれば，希望のものの複製を受け取ることが出来る．X は非公開利用 (EXCLUSIVE ACCESS) を意味し，このコードが付いたテキストは OUCS の登録者以外は使えない．これは東京大学大型計算機センターのサービスの条件と同じである。A は制限付きに利用 (RESTRICTED ACCESS) を意味し，このコードの付いたテキストは，利用者からの問い合わせは速やかに供託者に転送され，その許可のもとで OUCS は複製を作成する．又 OUCS の光学読み取り装置によって作成されたテキストは，12カ月間このカテゴリーが付された後 U のカテゴリーに移される．
2) 商品化されたテキスト・データベースには TLG 程の信頼度を持っているかどうかを宣伝文句に訳っている場合もある．
3) 富士 OCR, XP-50S, 富士電気総設．
4) E. G. Seidensticker "How Many People Wrote the Tale of Genji",「An Invitation to Japanese Literature」, 1974 財団法人日本文化研究所．
5) SGML 懇談会「報告書」, 1991.
6) OUCS のピーター・ロビンソンの開発したソフト "COLLATE" は，古代スカンジナビアのテキスト『Solarljod』の写本を 60 種類，ダンテの『De Monarchia』22 種類，チョーサーの『Wife of Bath's Prologue』では 58 種類の写本を同時に扱うことが出来る．